FENIX

RESEARCH INFRASTRUCTURE

D4.3

Infrastructure at JUELICH-JSC

Work package:	WP4 Procurement and deployment	
Author(s):	Dorian Krause	JUELICH
Reviewer #1	Dirk Pleiter	JUELICH
Reviewer #2	Thomas Leibovici	CEA
Dissemination Level	Public	
Nature	Other	

Date	Author	Comments	Version	Status
26.04.2020	Dorian Krause	First draft version.	V0.1	Draft
27.04.2020	Dirk Pleiter	Review	V0.2	Draft
27.04.2020	Thomas Leibovici	Review	V0.3	Draft
27.04.2020	Dorian Krause	Minor modifications to address reviewer comments	V1.0	Final

The ICEI project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreement No 800858.

Executive Summary

This document describes the infrastructure components operated by JUELICH-JSC for the Fenix infrastructure as of Q2 2020. It describes the hardware and software characteristics of these components and the status of the integration with other Fenix components.

Contents

Acronyms

AAI	Authentication and Authorization Infrastructure
ACD	Active Data Repositories
ACL	Access Control List
API	Application Programming Interface
ARD	Archival Data Repositories
BSC	Barcelona Supercomputing Center
CapEx	Capital Expenditure
CDP	Co-design Project
CEA	Commissariat à l'énergie atomique et aux énergies alternatives
CINECA	Consorzio Interuniversitario
CLI	Command Line Interface
CSCS	Centro Svizzero di Calcolo Scientifico
DL	Data Location Service
DM	Data Mover Service
DT	Data Transfer Service
FPA	Framework Partnership Agreement
FURMS	Fenix User and Resource Management Services
GoP	Group of Procurers
GUI	Graphical User Interface
HBP	Human Brain Project
HPAC	High Performance Analytics and Computing
HPC	High Performance Computing
HPDA	High Performance Data Analytics
HPST	High-Performance Storage Tier
IaaS	Infrastructure as a Service
IAC	Interactive Computing Services
ICCP	Interactive Computing Cloud Platform
ICEI	Interactive Computing E-Infrastructure for the Human Brain Project
ICN	Interactive Computing Node
IdP	Identity Provider
IPR	Intellectual Property Rights
JP	Joint Platform
JSC	Jülich Supercomputing Centre

LCST	Large-Capacity Storage Tier
MS	Monitoring Services
NDA	Non-Disclosure Agreement
NETE	External Interconnect
NETI	Internal Interconnect
NMC	Neuromorphic Computing
NVM	Non-Volatile Memory
NVRAM	Non-Volatile Random Access Memory
OIDC	OpenID Connect
OpEx	Operational Expenditure
PaaS	Platform as a Service
PCP	Pre-Commercial Procurement
PI	Principal Investigator
PID	Persistent Identifier
PIE	Public Information Event
PRACE	Partnership for Advanced Computing in Europe
Q&A	Questions and Answers
QoS	Quality of Service
R&D	Research & Development
R&I	Research & Innovation
RBAC	Role-Based Access Control
RFI	Request For Information
SCC	Scalable Computing Services
SGA	Specific Grant Agreement
SIB	Science & Infrastructure Board
SLA	Service Level Agreement
SP	Subproject
TCO	Total Cost of Ownership
TGCC	Très Grand Centre de calcul du CEA
UI	User Interface
US	User Support Services
VM	Virtual Machine Services

1. Introduction

In 2018, JUELICH-JSC started the procurement of infrastructure components for the ICEI project. Two major installations, the JUSUF compute and cloud platform and the JUST-IME high-performance storage systems, took place in late 2019 and early 2020. These infrastructure components enter their early operational phase soon. This document provides an overview of the components.

2. Overview of Infrastructure Components at JUELICH-JSC

JUELICH-JSC is starting operation of three systems that provide five of the Fenix services from the service catalogue. Table 1 summarizes these components, their correspondence with the services and provides an overview about the available resources.

				Quarterly allocation (2020)		
Component	**Service Type**	**ICEI Total Allocation**	**HBP Total Allocation**	**Q2**	**Q3**	**Q4**
JUSUF	SCC	186 nodes	43 nodes	75.542 node-hrs	75.542 node-hrs	75.542 node-hrs
	IAC	5 nodes	5 nodes	10.431 node-hrs	10.431 node-hrs	10.431 node-hrs
	VM	14 servers	4 servers	4 servers	4 servers	4 servers
JUST-IME	NVM	2 PB	0.5 PB	0.5 PB	0.5 PB	0.5 PB
JUST Object Storage	ARD	1 PB[1]	1 PB	1 PB	1 PB	1 PB

Table 1: Overview of the infrastructure components operated by JUELICH-JSC.

3. JUSUF (SCC, IAC, VM)

The JUSUF (Jülich Support for Fenix) system is a multi-purpose compute platform designed to offer three classes of services from the Fenix service catalogue on a single hardware platform: Scalable Computing, Interactive Computing and Virtual Machine Services. The system design enables JUELICH-JSC to adjust these shares of the resources assigned to each of the services about twice per year in accordance with the demand, cf. Figure 1.

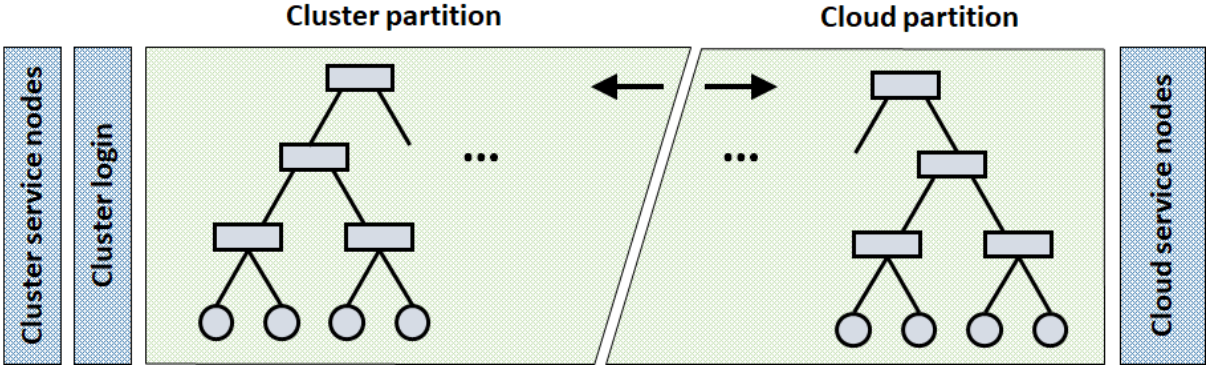

Figure 1: Sketch of the JUSUF system architecture.

[1] Resources are initially only available for the Human Brain Project.

JUSUF is a 205-node hybrid cluster/cloud resource delivered by Atos. All nodes are equipped with two latest-generation AMD multi-core processors (AMD EPYC "Rome" 7742 64-core processor, 2.2 GHz base frequency) and 256 GBytes main memory. All nodes are equipped with a dual-port ConnectX-6 adapter. Each node is connected to a HDR100 InfiniBand and a 40 Gbit/s Ethernet link. The system topology is a full fat tree. All nodes feature a local NVMe disk offering 1 TByte of fast user-accessible non-volatile memory with a file-system interface. 61 nodes are equipped with one NVidia V100 with 16 GBytes HBM2 memory. All nodes are connected with 40 Gb/s Ethernet to the Ethernet fabric in the JUELICH-JSC data centre. The four login nodes for the cluster partition are equipped with a 100 Gb/s Ethernet connection to external networks.

Figure 2: Photo of the JUSUF installation in the data centre. Please note that the rear-door heat exchangers have not yet been installed at the time the picture was taken.

More information for users is available on the systems' webpage [1].

Nodes can be either assigned to the cluster partition, serving Scalable Computing and Interactive Computing requirements, or the cloud partition, which provides the VM Service. The two partitions are strongly separated to account for the different security requirements of the high-performance computing and cloud services.

3.1 JUSUF Cluster Partition (SCC, IAC)

The JUSUF cluster partition provides batch and interactive computing capabilities. The system is installed with the most recent CentOS 7 version and uses the Slurm workload manager in conjunction with the ParaStation resource management system. The same system software stack is currently available on the JUWELS and JURECA supercomputers operated by JUELICH-JSC. Optimizing compilers and several MPI implementations are provided along with installations of commonly used software tools managed by JUELICH-JSC.

The nodes in the cluster partition use the 40 Gb/s Ethernet connection to the facility Ethernet fabric to access the high-performance file systems (ACDs), based on IBM Spectrum Scale, provided by JUST (Jülich Storage Cluster). The referenced subsystem of the JUST storage complex, provides a raw capacity of 70 PBytes with a file system bandwidth of up to 450 GBytes/s to different client systems, including the Tier-0/1 system JUWELS, in the JUELICH-JSC facility. Hence, JUSUF users with other allocations on a JSC machine can access the same storage. In addition, users can leverage the node-local NVMe storage and the JUST-IME service (see Section 4) for I/O acceleration.

SCC and IAC Services are mapped to different Slurm partitions with appropriate scheduling configurations and policies. The optimization of these settings are subject to future use-case driven investigations in conjunction with the work in WP6 of the Human Brain Project.

The system is accessible via four dedicated login nodes via SSH. In addition, UNICORE-based access and support for the JUELICH-JSC JupyterHub are currently made available.

The SCC and IAC Services on the JUSUF cluster partition will be available starting first of May, when it enters the early operational phase.

3.2 JUSUF Cloud Partition (VM)

The cloud partition of the JUSUF system provides Infrastructure-as-a-Service (IaaS) capabilities based on the Red Hat OpenStack Platform (RHOSP). At the time of this writing RHOSP version 16, based on Red Hat Enterprise Linux 8, is installed. The OpenStack installation uses the 40 Gbits/s Ethernet connection for project-internal and external network connectivity (the latter via router nodes).

The VM Service on JUSUF supports all VM types specified in the Fenix VM Model. The local 1 TByte NVMe drive can be made accessible to virtual machines. Support for GPU sharing with the NVidia vGPU feature has been tested and is available for users, albeit in the state of a technical preview.

The integration with the recently established first-generation of the Fenix AAI is on-going. Initially, it is expected that only users of the web-interfaces (OpenStack Horizon) will be able to authenticate with the Fenix AAI, while command-line access to other APIs will not work with the Fenix AAI.

Access to the VM Service will be opened up before End of May.

4. JUST-IME High-Performance Storage (ACD, NVM)

The JUST (Jülich Storage) system [2] is the central storage provider in JUELICH-JSC's supercomputer facility and provides different storage services, from high-performance storage to long-term archiving services.

For the ICEI project a new storage service was developed and deployed to provide a bandwidth-optimized, globally accessible and consistent, storage layer suitable also for random access I/O. This new service, JUST-IME, is based on the DataDirect Network (DDN) "Infinite Memory Engine" (IME)

software and IME-140 storage appliances. The system is delivered by Hewlett-Packard Enterprise together with DDN. It consists of a total of 110 IME-140 servers with an accumulated capacity of ca. 2 PBytes and a nominal bandwidth of more than 2 TBytes/s, accumulatively available to the JURECA Cluster, JUWELS and JUSUF installations in Jülich. The sustained bandwidth in practice is slightly lower due to routing effects. In contrast to other JUST services, which are made available to client systems via JUELICH-JSC's facility Ethernet fabric, the JUST-IME service is directly integrated into the high-speed InfiniBand-based interconnects of the client systems and utilizes remote direct memory access semantics for improved performance.

A unique feature of the JUST-IME system is a slice-based architecture (cf. Figure 3), which allows the system to be integrated with multiple client systems, each with a separate high-speed interconnect fabric, while still maintaining a globally consistent namespace. Writes from each system will always be handled by the local slice but the data can be read from a different client system with only small performance implications. This capability is not yet available in the initial deployment in Q2 2020 but will be made available through a series of software updates until 2021. Currently each slice is an individual storage system serving only its directly attached client compute system.

The system is split into three slices according to the client systems JURECA Cluster, JUWELS and JUSUF. The slice shares are approximately 40%/50%/10%. The capacity of the JUSUF slice equals approximately 220 TBytes.

Users can request access to the JUST-IME NVM Service in conjunction with an ICEI allocation on JUSUF and/or a PRACE or GCS compute time grant on the JUWELS system. Users with a compute time allocation on the JURECA Cluster may also request access to the JURECA JUST-IME slice via ICEI.

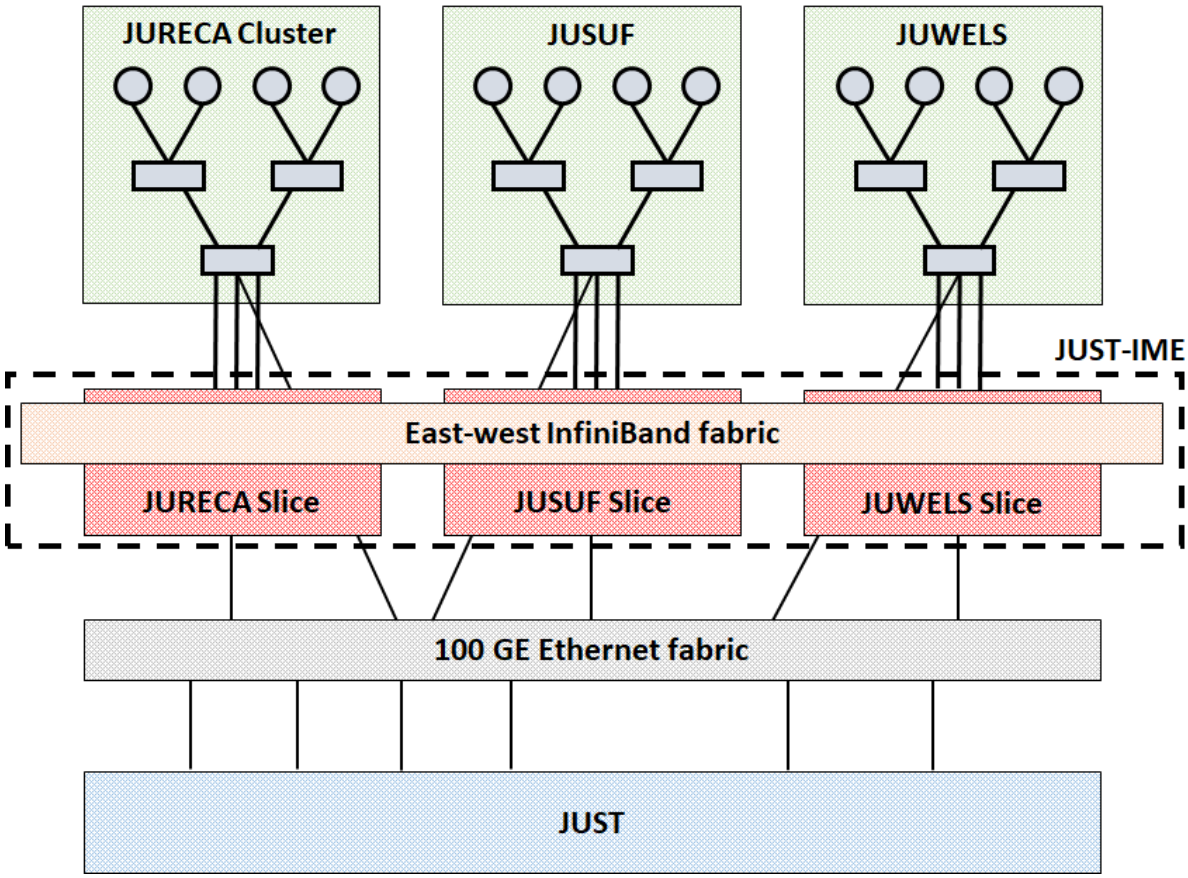

Figure 3: Sketch of the JUST-IME system architecture.

The system enters the early operational phase in May. Due to on-going tests, the JUWELS slice will be regularly in maintenance during the first weeks.

5. JUST Object Storage (ARC)

The archival data repository (ARC) Service is implemented by the JUST object storage service. It is implemented with the IBM Spectrum Scale Cluster Export Services (CES) feature on top of a capacity-optimized hardware platform that also serves an active data repository to the facility.

The ARC Service is currently in an early operational/pilot phase and available for interested users. The integration with the Fenix AAI will be performed once the necessary prerequisites by the AAI are met.

Up to 1 PByte of storage is available for users of the Human Brain Project. JUELICH-JSC currently does not offer ARD storage resources for other ICEI users as the storage is not funded through the ICEI project.

6. Concluding remarks

This document provides an overview of the ICEI resources available at JUELICH-JSC starting in Q2 2020. The described infrastructure components are currently or enter soon in their early operational phase and are available to interested users.

7. References

[1] http://fz-juelich.de/ias/jsc/jusuf

[2] Jülich Supercomputing Centre. (2019. JUST: Large-Scale Multi-Tier Storage Infrastructure at the Jülich Supercomputing Centre. Journal of large-scale research facilities, 5, A136. http://dx.doi.org/10.17815/jlsrf-5-172